

RGB–wearable sensor fusion improves isolated signlanguage recognition under visually challenging conditions

Xinao Han

Tianjin University of Technology

Chengjie Huang

Tianjin University of Technology

Wei Wang

wwang@email.tjut.edu.cn

Tianjin University of Technology

Tian Xie

Tianjin University of Technology

Zhuonan Li

Hebei University of Technology

Chi Zhang

Tianjin University of Technology

Yu Song

Tianjin University of Technology

Junchao Zhu

Tianjin University of Technology

Article

Keywords:

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10280059/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

RGB–wearable sensor fusion improves isolated sign language recognition under visually challenging conditions

Xinao Han^{1,2}, Chengjie Huang^{1,2}, Wei Wang^{1,2,*}, Tian Xie^{1,2}, Zhuonan Li^{3,4}, Chi Zhang^{1,2}, Yu Song^{1,2}, and Junchao Zhu^{1,2}

¹School of Electrical Engineering and Automation, Tianjin University of Technology, Tianjin 300384, China

²Tianjin Key Laboratory of New Energy Power Conversion, Transmission and Intelligent Control, Tianjin 300384, China

³School of Mechanical Engineering, Hebei University of Technology, Tianjin, China

⁴Tianjin Zwinsoft Technology Co., Ltd., Tianjin, China

*Corresponding author: wwang@email.tjut.edu.cn

ABSTRACT

RGB-based sign language recognition (SLR) is sensitive to motion blur, hand occlusion, illumination variation and background clutter, which can weaken the visual evidence needed for fine-grained sign discrimination. Here, we present RGB–Wearable Sensor Fusion (RWS-Fusion), a multimodal framework for isolated SLR that combines RGB video with synchronized wearable sensing. The RGB branch uses a three-dimensional Swin Transformer-Tiny (Swin3D-T) backbone to extract spatiotemporal visual representations from uniformly sampled video frames. The wearable branch integrates flex sensors and multi-site inertial measurement units to capture finger flexion, hand motion, wrist posture and upper-limb dynamics. Normalized sensor streams are augmented with first-order temporal differences and encoded using a lightweight Sensor Transformer. For multimodal inference, modality-specific temperature scaling is applied to RGB and sensor logits, and validation-selected weighted probability fusion is used to generate the final prediction. Experiments on a synchronized RGB–wearable dataset comprising 5 participants, 100 isolated sign classes and 5,000 aligned samples were conducted under leave-one-subject-out evaluation. RWS-Fusion achieved 97.10% accuracy and 96.83% macro F1-score, outperforming RGB-only and sensor-only baselines. These findings suggest that wearable motion cues provide complementary information to RGB video for isolated SLR in visually degraded acquisition conditions.

Introduction

Sign language recognition (SLR) aims to automatically interpret signed gestures from sensory data and has applications in assistive communication and human–computer interaction. Recent work has investigated both vision-based sign recognition and wearable-sensor gesture decoding, including hybrid deep networks for sign classification, RGB-based modelling of manual and non-manual cues, and low-cost data-glove sensing^{1–4}. A recent survey indicates that SLR research has moved from hand-crafted features towards deep spatiotemporal representation learning⁵.

Large-scale datasets and benchmark studies have supported more systematic evaluation. The Word-Level American Sign Language dataset, AUTSL, Sign Language Transformers and How2Sign enable evaluation across larger vocabularies, signers and acquisition settings^{6–9}. General video backbones, including Video Swin Transformer, Video Vision Transformer, Video Masked Autoencoder and Multiscale Vision Transformer v2, have also improved RGB video representation learning^{10–13}. SLR-specific methods further enhance hand-aware representation, temporal alignment and continuous sign modelling through SignBERT, visual alignment constraints, consistency-enhanced continuous SLR, correlation networks and contrastive visual–textual transformation^{14–18}.

Despite these advances, RGB-only SLR remains vulnerable when visual evidence is degraded. Motion blur, hand occlusion, illumination variation, background clutter, low-quality frames and partial hand truncation can reduce the reliability of hand appearance and motion trajectory cues. Skeleton- and keypoint-based methods can reduce background interference, but they still depend on accurate landmark extraction, which may fail when hand boundaries, fingertip locations or body-reference directions are visually ambiguous¹⁹.

Wearable sensors provide complementary motion information. Flex sensors, inertial measurement units (IMUs), data gloves and other wearable sensors can directly capture finger flexion, hand orientation, acceleration, angular velocity and upper-limb dynamics^{20–23}. Smart-glove systems have also demonstrated the potential of wearable sensing for sign interpretation and

human-machine interaction^{24,25}. However, sensor-only recognition can confuse signs with similar finger-bending states or inertial trajectories, and it does not explicitly encode visual appearance, spatial context or body-referenced direction.

Here, we present RGB-Wearable Sensor Fusion (RWS-Fusion), an RGB-wearable sensor fusion framework for isolated SLR under controlled acquisition conditions that include common visual degradations. The study makes three main contributions. First, we construct a synchronized RGB-wearable sensor dataset for isolated SLR, containing 5 participants, 100 sign classes and 5,000 aligned multimodal samples. Second, we develop a multi-stream wearable sensor branch for temporal motion modelling. Flex sensors and multi-site IMUs capture finger flexion, hand motion, wrist posture and upper-limb dynamics, while first-order temporal differences and a Lite Sensor Transformer model motion states and temporal variations. Third, we introduce a lightweight temperature-calibrated decision-level fusion strategy. Modality-specific temperature calibration is applied to RGB and wearable sensor logits, and calibrated class probabilities are fused using a validation-optimized weight. Experiments show that RWS-Fusion improves performance over RGB-only and sensor-only baselines under leave-one-subject-out evaluation.

Results

RGB-only recognition under cross-subject evaluation

Four video recognition models, R3D, VideoMAE, MVITv2-S and Swin3D-T, were evaluated using the same leave-one-subject-out protocol¹⁰⁻¹³. As summarized in Table 1, Swin3D-T achieved the highest RGB-only performance, with 67.54% accuracy, 61.43% macro F1-score and 85.74% Top-5 accuracy. RGB-only recognition nevertheless remained limited under cross-subject evaluation, consistent with the sensitivity of visual models to hand occlusion, illumination variation, motion blur, low-quality frames and partial hand truncation.

Wearable sensor recognition under cross-subject evaluation

Several temporal models were evaluated using the same wearable sensor inputs, including 1D-CNN, BiLSTM, CNN-BiLSTM, Transformer and the proposed Lite Sensor Transformer²⁶⁻²⁸. As summarized in Table 1, the Lite Sensor Transformer achieved the highest sensor-only performance, with 95.20% accuracy, 94.61% macro F1-score and 99.04% Top-5 accuracy. It improved accuracy and macro F1-score by 10.12 and 10.75 percentage points over 1D-CNN, and by 4.34 and 4.73 percentage points over CNN-BiLSTM, respectively. These results indicate that multi-site wearable sensing, temporal-difference enhancement and lightweight Transformer-based temporal encoding are suitable for heterogeneous wearable time-series signals.

Decision-level fusion performance

After the modality-specific branches were evaluated, Swin3D-T and the Lite Sensor Transformer were selected as the RGB and wearable-sensor backbones, respectively. RGB-only and Sensor-only served as single-modal baselines, and several decision-level fusion variants were compared with RWS-Fusion under the same cross-subject protocol²⁹⁻³². Average Fusion combined RGB and sensor softmax probabilities with fixed equal weights. Alpha-only Fusion optimized a global modality weight without temperature calibration. Temperature-only Fusion applied modality-specific temperature calibration with equal fusion weights. Confidence-adaptive Fusion adjusted modality weights according to prediction confidence. RWS-Fusion combined modality-specific temperature calibration with validation-optimized global weighting.

As summarized in Table 1, Sensor-only substantially outperformed RGB-only, increasing accuracy from 67.54% to 95.20% and macro F1-score from 61.43% to 94.61%. This result shows that wearable sensor signals provided more stable and discriminative motion cues than RGB video in the present setting. All fusion variants further improved over the single-modal baselines, supporting the complementarity between RGB video and wearable sensor signals. RWS-Fusion achieved the highest overall performance, with 97.10% accuracy, 97.62% precision, 96.83% macro F1-score and 99.68% Top-5 accuracy. The improvement over the sensor-only baseline was moderate but corresponded to a reduction in misclassified samples, as shown in Fig. 1.

Class-wise behaviour and error analysis

To examine class-wise recognition behaviour, Fig. 1 shows the 100-class row-normalized confusion matrices for RGB-only, Sensor-only and RWS-Fusion under cross-subject evaluation. The RGB-only model produced a scattered off-diagonal distribution, with 1,623 misclassified samples and 67.54% accuracy. Sensor-only recognition showed a clearer diagonal structure and reduced the number of errors to 240, corresponding to 95.20% accuracy. RWS-Fusion further reduced the number of errors to 145 and achieved 97.10% accuracy, yielding the most concentrated diagonal distribution among the three settings.

Fig. 2 summarizes fusion behaviour under RGB-sensor agreement and disagreement. When both unimodal branches were correct, all 3,206 samples were also correctly classified after fusion, indicating that RWS-Fusion did not introduce errors when both modalities were reliable. When the RGB branch was incorrect but the sensor branch was correct, fusion correctly classified 1,515 samples and misclassified 39 samples. When the RGB branch was correct but the sensor branch was incorrect, fusion correctly classified 130 samples and misclassified 41 samples. When both branches were incorrect, 4 samples were corrected

Table 1. Model comparison under leave-one-subject-out evaluation. Values are mean $\pm$ standard deviation across five folds.

Setting	Model or method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Top-5 accuracy (%)
RGB visual model	R3D	65.34 $\pm$ 11.39	59.76 $\pm$ 14.38	65.34 $\pm$ 11.39	58.53 $\pm$ 13.38	84.97 $\pm$ 13.24
	VideoMAE	62.31 $\pm$ 8.40	57.80 $\pm$ 12.21	62.31 $\pm$ 8.40	56.10 $\pm$ 10.29	84.08 $\pm$ 12.11
	MViTv2-S	66.29 $\pm$ 9.53	61.36 $\pm$ 13.70	66.29 $\pm$ 9.53	59.94 $\pm$ 11.72	85.42 $\pm$ 13.16
	Swin3D-T	67.54 $\pm$ 9.25	63.03 $\pm$ 12.37	67.54 $\pm$ 9.25	61.43 $\pm$ 11.16	85.74 $\pm$ 12.07
Wearable sensor model	1D-CNN	85.08 $\pm$ 2.76	87.65 $\pm$ 3.42	85.08 $\pm$ 2.76	83.86 $\pm$ 3.56	97.50 $\pm$ 1.48
	BiLSTM	92.58 $\pm$ 4.49	92.83 $\pm$ 4.60	92.58 $\pm$ 4.49	91.76 $\pm$ 5.14	98.76 $\pm$ 1.30
	CNN-BiLSTM	90.86 $\pm$ 4.68	91.93 $\pm$ 4.55	90.86 $\pm$ 4.68	89.88 $\pm$ 5.39	98.68 $\pm$ 0.86
	Transformer	92.92 $\pm$ 4.53	92.69 $\pm$ 5.17	92.92 $\pm$ 4.53	91.92 $\pm$ 5.44	98.84 $\pm$ 0.80
	Lite Sensor Transformer	95.20 $\pm$ 2.43	95.07 $\pm$ 2.88	95.20 $\pm$ 2.43	94.61 $\pm$ 2.94	99.04 $\pm$ 1.11
Fusion strategy	RGB-only	67.54 $\pm$ 9.25	63.03 $\pm$ 12.37	67.54 $\pm$ 9.25	61.43 $\pm$ 11.16	85.74 $\pm$ 12.07
	Sensor-only	95.20 $\pm$ 2.43	95.07 $\pm$ 2.88	95.20 $\pm$ 2.43	94.61 $\pm$ 2.94	99.04 $\pm$ 1.11
	Average Fusion	96.42 $\pm$ 1.76	96.78 $\pm$ 1.67	96.42 $\pm$ 1.76	95.98 $\pm$ 2.16	99.62 $\pm$ 0.38
	Alpha-only Fusion	96.00 $\pm$ 2.29	96.66 $\pm$ 1.75	96.00 $\pm$ 2.29	95.46 $\pm$ 2.84	99.64 $\pm$ 0.38
	Temperature-only Fusion	96.36 $\pm$ 1.38	97.09 $\pm$ 1.03	96.36 $\pm$ 1.38	95.90 $\pm$ 1.74	99.68 $\pm$ 0.41
	Confidence-adaptive Fusion	96.28 $\pm$ 1.71	96.68 $\pm$ 1.56	96.28 $\pm$ 1.71	95.85 $\pm$ 2.10	99.64 $\pm$ 0.43
	RWS-Fusion	97.10 $\pm$ 1.50	97.62 $\pm$ 1.14	97.10 $\pm$ 1.50	96.83 $\pm$ 1.72	99.68 $\pm$ 0.41

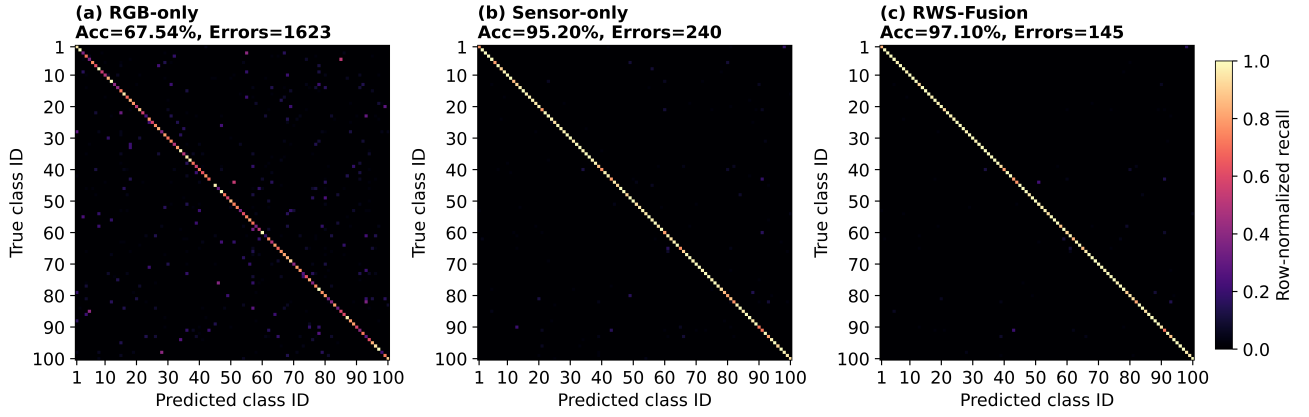


Figure 1. Full 100-class row-normalized confusion matrices under cross-subject evaluation. (a) RGB-only. (b) Sensor-only. (c) RWS-Fusion.

by fusion and 65 remained misclassified. These results indicate that RWS-Fusion generally exploits the more reliable modality in disagreement cases, whereas samples ambiguous for both modalities remain challenging.

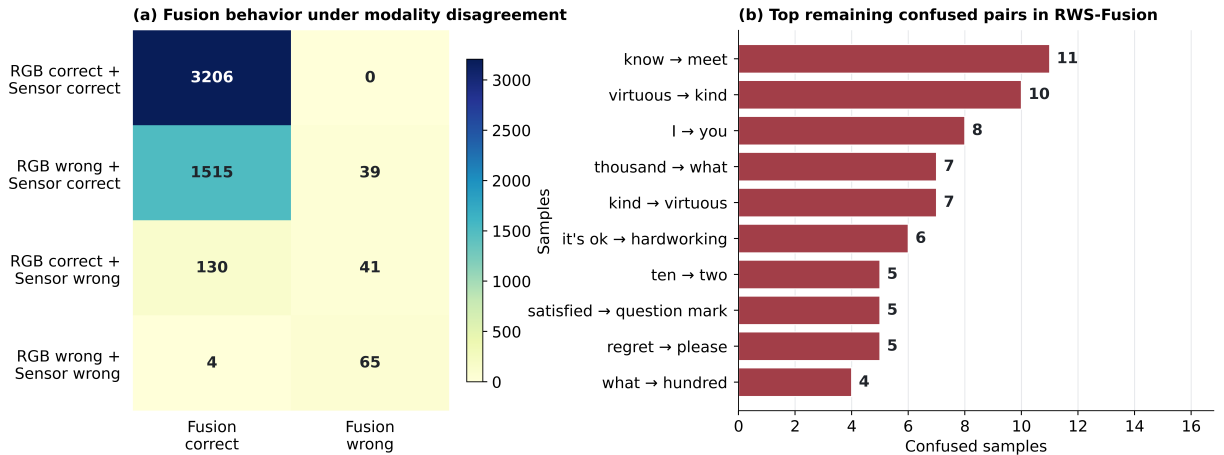


Figure 2. Error analysis of RWS-Fusion under cross-subject evaluation. (a) Fusion behavior under modality agreement and disagreement. (b) Top remaining confused class pairs in RWS-Fusion.

The main residual confusions in RWS-Fusion involved pairs such as “know” and “meet”, “virtuous” and “kind”, “I” and “you”, and “thousand” and “what”. These errors are consistent with similar hand shapes, close motion trajectories, subtle pointing directions and body-reference ambiguity^{33–36}. For example, distinguishing “I” from “you” depends on fine-grained

pointing direction, whereas distinguishing “virtuous” from “kind” may require more discriminative modelling of hand posture and finger-flexion patterns. Overall, residual errors were concentrated in a small number of visually or kinematically similar sign pairs, suggesting that future work should focus on finer modelling of hand orientation, body-reference direction and subtle temporal differences. The model architecture, acquisition workflow and implementation settings are described in the Methods section.

Discussion

This study presents RWS-Fusion, a multimodal framework for isolated SLR that combines RGB video, wearable sensing and temperature-calibrated decision-level fusion. Under leave-one-subject-out evaluation on a synchronized RGB–wearable dataset, RWS-Fusion achieved 97.10% accuracy, 96.83% macro F1-score and 99.68% Top-5 accuracy, outperforming both RGB-only and sensor-only baselines.

The results show that wearable sensing provides robust motion information when RGB evidence is degraded by occlusion, motion blur, illumination variation or background clutter. The sensor-only branch achieved strong standalone performance, but RGB information remained useful because it encodes visual appearance and body-referenced spatial cues that are not directly captured by inertial and flex sensors. The reduction from 240 to 145 errors indicates that calibrated decision-level fusion can correct a subset of modality-specific failures, particularly when RGB and wearable sensors provide complementary evidence.

Several limitations should be considered. The dataset includes 5 participants and should be expanded to include more signers, signer styles, acquisition environments, lighting conditions, backgrounds and wearing conditions to further evaluate generalization^{37,38}. The present study focuses on isolated sign recognition; continuous SLR additionally requires temporal segmentation, co-articulation modelling and language-level understanding^{8,9,16,17}. Future work will investigate larger-scale data collection, signer-independent calibration, adaptive fusion strategies and real-time deployment of RGB–wearable sensor fusion systems^{39–44}.

Methods

Overview of RWS-Fusion

Given a sign language sample, the RGB video is represented as a sequence of RGB frames, and the wearable sensor modality is represented as three wearable sensor streams:

$$V = \{I_1, I_2, \dots, I_T\}, \quad S = \{S_b, S_h, S_f\}, \quad (1)$$

where I_t denotes the t -th RGB frame, T is the number of frames in the original video, and S_b , S_h and S_f denote the body IMU signal, hand IMU signal and hand flex signal, respectively.

RWS-Fusion processes the RGB and wearable modalities using two modality-specific branches, as shown in Fig. 3. The RGB branch captures hand appearance, body posture, motion trajectory and spatial context from uniformly sampled frames, whereas the wearable branch models body IMU, hand IMU and hand flex signals. Both branches first produce class logits over the same sign-class space, which are then temperature-calibrated, converted into class probabilities and combined using a validation-selected fusion weight.

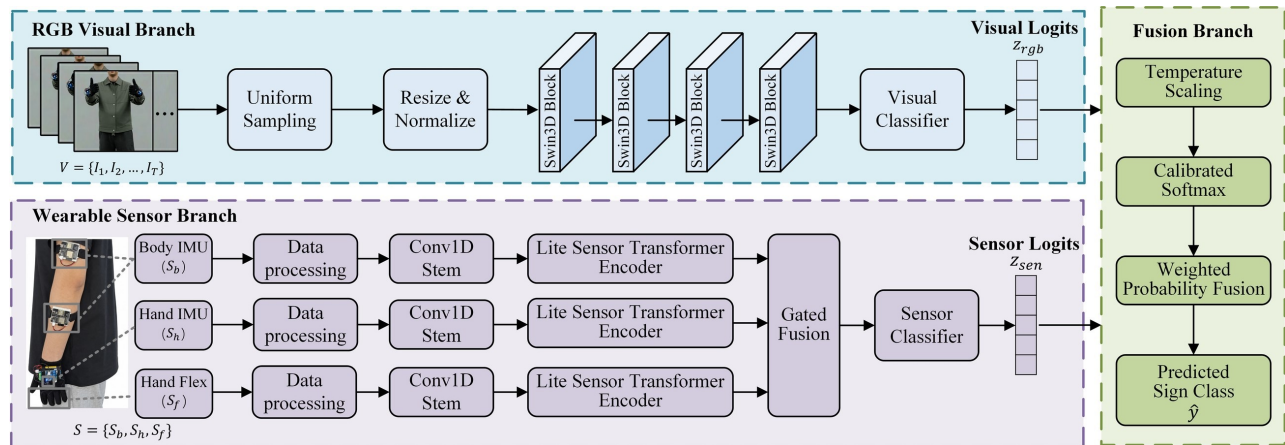


Figure 3. Overall architecture of the proposed RWS-Fusion framework.

Dataset and acquisition system

A synchronized RGB–wearable acquisition system was developed to construct the multimodal SLR dataset, as shown in Fig. 4^{20,23,24}. The system comprised a USB camera, a PC acquisition unit, a receiver ESP32 module and three wearable sensing nodes positioned on the upper arm, forearm and hand. The wearable hardware included ESP32-S3-WROOM-1 boards, ESP32-C3 boards, JY901S IMU modules, batteries and flex sensors. The upper-arm and forearm nodes recorded upper-limb inertial motion, whereas the hand node captured local hand movement, wrist posture and finger flexion.

During data collection, the wearable nodes transmitted sensor packets to the receiver ESP32 module through ESP-NOW communication. The receiver forwarded multi-node sensor data to the PC through a USB serial interface. In parallel, the RGB camera streamed video to the same PC through a USB camera interface. The PC served as the central acquisition unit and stored RGB frames and wearable sensor signals in a synchronized acquisition process.

The acquisition design follows the same modality structure as the recognition model. The RGB camera preserves visual appearance and body-referenced spatial information, whereas the upper-limb IMU, hand IMU and flex sensors provide synchronized physical motion cues. These streams were stored together but remained separable for model training, enabling the RGB and wearable branches to be optimized independently before decision-level fusion.

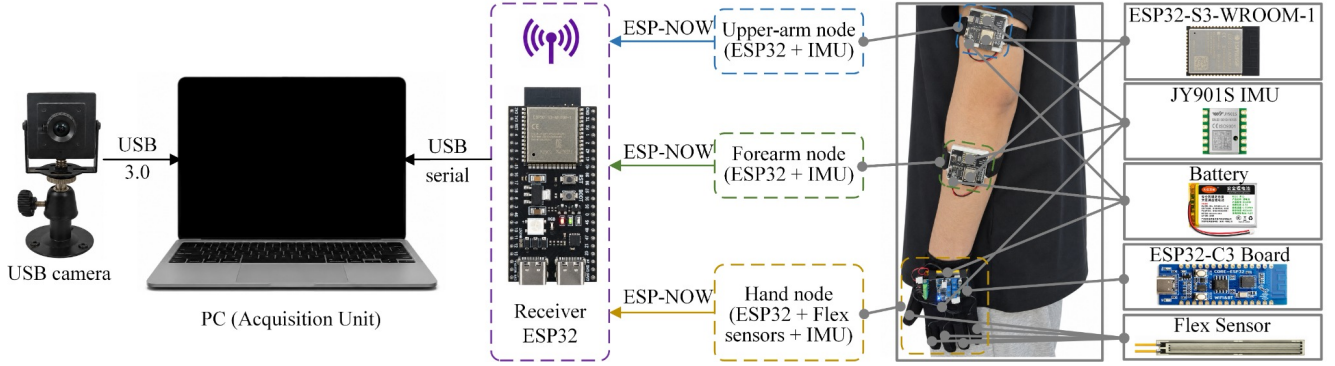


Figure 4. Integrated RGB–wearable acquisition system.

The dataset contains 5 participants, 100 isolated sign classes and 5,000 aligned RGB–wearable samples. Each participant performed each sign class 10 times, resulting in 1,000 samples per participant. Each sample consists of a time-aligned RGB video and synchronized wearable sensor streams. The visually challenging acquisition setting included common degradations such as illumination variation, partial hand occlusion, motion blur and background clutter while keeping the sign vocabulary and acquisition protocol consistent across participants. The RGB-only, sensor-only and fusion comparisons are summarized in Table 1.

Fig. 5 illustrates a representative aligned sample, including uniformly sampled RGB frames, ten-channel finger-bending voltage curves and a motion-summary heatmap derived from six nine-axis IMUs.

Ethics approval and consent to participate

This study involved non-invasive RGB video recording and wearable sensor data collection from adult volunteers in a controlled laboratory setting. The study protocol was reviewed and approved by the relevant ethics review body of Tianjin University of Technology (approval no. TUT-R-20260313-007). All procedures were performed in accordance with relevant institutional guidelines and regulations. All participants were informed of the study purpose and data collection procedure, and written informed consent was obtained before data collection. Written informed consent was also obtained for the publication of identifiable images and related data in an online open-access journal.

RGB visual branch

For each RGB video, 32 frames were uniformly sampled from the original sequence. If a video contained fewer than 32 frames, the last frame was repeated until the required length was reached. Each sampled frame was resized to 224×224 , normalized and used as input to the RGB branch.

The RGB encoder was implemented using Swin3D-T, following the Video Swin Transformer architecture¹⁰. The model extracts hierarchical spatiotemporal features from the sampled RGB sequence and outputs visual logits through a classification head composed of dropout and a linear layer:

$$z_{\text{rgb}} = C_{\text{rgb}}(B_{\text{swin}}(X_{\text{rgb}})), \quad (2)$$

where $B_{\text{swin}}(\cdot)$ denotes the Swin3D-T backbone, $C_{\text{rgb}}(\cdot)$ denotes the RGB classification head, and z_{rgb} represents the visual logits over sign classes.

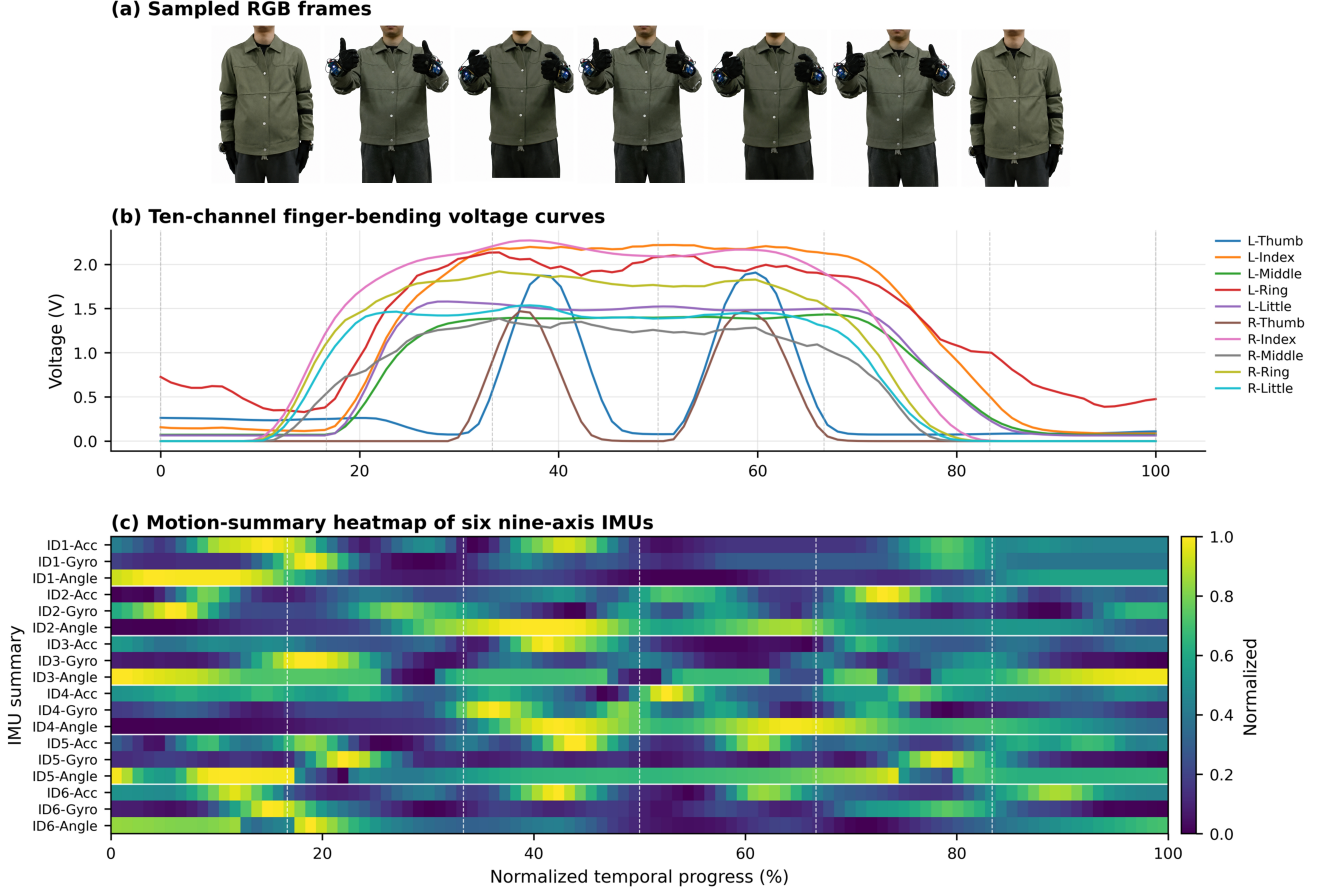


Figure 5. Representative aligned RGB-wearable sample.

Wearable sensor branch

The wearable sensor branch models sign motion from three synchronized sensor groups: body IMU, hand IMU and hand flex^{20–23}. The body IMU stream describes upper-limb inertial motion, the hand IMU stream captures local hand motion and wrist posture, and the hand flex stream records finger-bending states.

To reduce the influence of different numerical ranges among sensor channels, each channel is normalized using z-score normalization:

$$\hat{s}_{m,l,c} = \frac{s_{m,l,c} - \mu_{m,c}}{\sigma_{m,c} + \epsilon}, \quad (3)$$

where $s_{m,l,c}$ denotes the value of the c -th channel in the m -th sensor stream at the l -th time step, $m \in \{b, h, f\}$, $\mu_{m,c}$ and $\sigma_{m,c}$ denote the mean and standard deviation of the corresponding channel, and ϵ is a small constant used to avoid division by zero.

To enhance temporal motion modelling, the first-order temporal difference is calculated as:

$$\Delta \hat{s}_{m,l,c} = \hat{s}_{m,l,c} - \hat{s}_{m,l-1,c}. \quad (4)$$

The first time step is padded with zero to keep the sequence length unchanged. The normalized signal and its temporal difference are concatenated along the channel dimension:

$$X_m = \text{Concat}(\hat{S}_m, \Delta \hat{S}_m), \quad (5)$$

where X_m denotes the enhanced input of the m -th sensor stream. This design allows the model to use both motion states and temporal variations, such as finger-bending changes, wrist rotation and inertial transitions.

Each enhanced sensor stream is processed by a Conv1D stem and a Lite Sensor Transformer Encoder, as shown in Fig. 6. The Conv1D stem consists of a temporal convolution, batch normalization, GELU activation, a depthwise temporal convolution and dropout. It projects each stream into a 128-dimensional embedding space and reduces the temporal length before Transformer encoding.

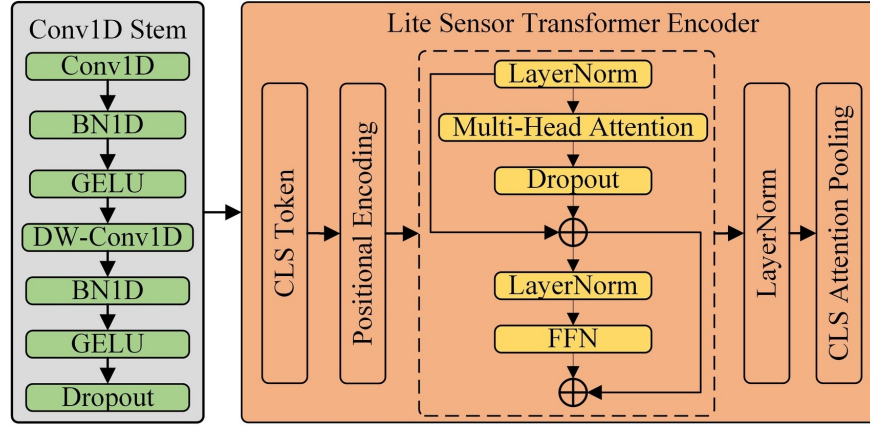


Figure 6. Lite Sensor Transformer Encoder used for each wearable sensor stream.

The Lite Sensor Transformer Encoder contains 3 Transformer blocks with 4 attention heads and an MLP expansion ratio of 2.0. A learnable CLS token and positional encoding are added to preserve temporal structure. After Transformer encoding, the stream-level feature is obtained as:

$$f_m = E_m(X_m), \quad m \in \{b, h, f\}, \quad (6)$$

where $E_m(\cdot)$ denotes the corresponding Lite Sensor Transformer Encoder and f_m denotes the encoded feature of the m -th sensor stream.

To integrate the three sensor streams, a gated fusion module is used. The three stream-level features are first concatenated, and the gating weights are computed as:

$$g = \text{Softmax} \left(W_2 \delta \left(W_1 [f_b; f_h; f_f] + b_1 \right) + b_2 \right), \quad (7)$$

where $g = [g_b, g_h, g_f]$ denotes the learned weights for the body IMU, hand IMU and hand flex streams, respectively. W_1 , W_2 , b_1 and b_2 are learnable parameters of the gate network, $\delta(\cdot)$ denotes the GELU activation, and $[f_b; f_h; f_f]$ denotes feature concatenation.

The weighted stream features are concatenated and projected to obtain the fused sensor representation:

$$f_{\text{sen}} = P \left([g_b f_b; g_h f_h; g_f f_f] \right), \quad (8)$$

where $P(\cdot)$ denotes the projection network composed of a linear layer, GELU activation, dropout and another linear layer. The fused representation f_{sen} is fed into a classifier composed of layer normalization, dropout and a linear layer to generate the sensor logits z_{sen} .

Temperature-calibrated decision-level fusion

RGB videos and wearable sensor signals provide complementary information, but their logits may have different confidence distributions^{45,46}. Therefore, RWS-Fusion adopts a lightweight temperature-calibrated decision-level fusion strategy. Instead of fusing heterogeneous intermediate features, the proposed method calibrates and combines the output logits of the RGB and wearable sensor branches.

Let z_{rgb} and z_{sen} denote the logits generated by the RGB branch and the wearable sensor branch, respectively. Temperature scaling is first applied to each modality:

$$p_m = \text{softmax} \left(\frac{z_m}{T_m} \right), \quad m \in \{\text{rgb}, \text{sen}\}, \quad (9)$$

where p_m is the calibrated probability vector of modality m and T_m is the corresponding temperature parameter. The calibrated probability vectors are then fused as:

$$p_f = \alpha p_{\text{rgb}} + (1 - \alpha) p_{\text{sen}}, \quad (10)$$

where p_f denotes the fused probability vector, α is the fusion weight assigned to the RGB branch, and $1 - \alpha$ is the fusion weight assigned to the wearable sensor branch. The final predicted class is obtained as:

$$\hat{y} = \arg \max_c p_{f,c}, \quad (11)$$

where $p_{f,c}$ denotes the fused probability of the c -th class.

Evaluation protocol

All experiments used leave-one-subject-out cross-subject evaluation. In each fold, one participant was held out as the test subject, and the remaining four participants were used for model training and validation. The validation subset was selected only from the training subjects and was used for model selection, temperature calibration and fusion-weight selection. The held-out test subject was not used for hyperparameter selection, temperature fitting or fusion-weight optimization.

The RGB branch, wearable sensor branch and fusion variants were evaluated using the same train–validation–test split in each fold. The trained RGB and sensor models were fixed before decision-level fusion, and the validation subset was used only for temperature selection and fusion-weight selection. All reported values are the mean $\pm$ standard deviation across the five leave-one-subject-out folds.

Implementation details

All experiments were implemented in PyTorch and run on a single NVIDIA GeForce RTX 3090 GPU. The RGB and wearable sensor models were trained independently, and fusion was performed only after single-modal training.

For the RGB branch, each video was resized to 224×224 pixels and normalized before input to Swin3D-T initialized with Kinetics-400 pretrained weights. The classification head was replaced for 100-class classification. The model was trained for up to 50 epochs using cross-entropy loss with label smoothing of 0.1, AdamW optimization, a batch size of 6 and a weight decay of 5×10^{-2} . The backbone was frozen for the first 3 epochs with a head learning rate of 8×10^{-4} , and then fine-tuned with learning rates of 5×10^{-6} for the backbone and 5×10^{-5} for the classification head. Early stopping with a patience of 6 epochs was applied based on validation accuracy.

For the wearable sensor branch, body IMU, hand IMU and hand flex streams were temporally standardized, normalized using training-fold statistics, and concatenated with first-order temporal differences. Each stream was encoded by a Conv1D stem and a Lite Sensor Transformer with 128-dimensional embeddings, 3 Transformer blocks and 4 attention heads, followed by gated sensor-stream fusion. The model was trained for up to 50 epochs using focal cross-entropy loss with $\gamma = 2.0$, label smoothing of 0.05, AdamW optimization, a batch size of 512, a learning rate of 1×10^{-3} and a weight decay of 1×10^{-4} . Early stopping with a patience of 6 epochs was applied based on validation macro F1-score.

For decision-level fusion, the trained RGB and sensor models were fixed and used to generate logits for aligned multimodal samples. The RGB and sensor temperature values were selected from $\{0.6, 0.7, 0.8, 1.0, 1.2, 1.5, 2.0\}$, and the RGB fusion weight α was selected from 0.05 to 0.95 with a step size of 0.05. The fused probability was computed as $\alpha p_{\text{rgb}} + (1 - \alpha) p_{\text{sen}}$. The temperature values and fusion weight were selected on the validation subset by maximizing macro F1-score and were then fixed for evaluation on the held-out test subject. Accuracy, macro precision, macro recall, macro F1-score and Top-5 accuracy were reported as mean $\pm$ standard deviation across five folds.

Conclusion

This study developed RWS-Fusion, an RGB–wearable sensor fusion framework for isolated sign language recognition under visually challenging acquisition conditions. The framework combines Swin3D-T-based RGB representation learning, multi-stream wearable sensing and temperature-calibrated weighted probability fusion. Under leave-one-subject-out evaluation, RWS-Fusion achieved 97.10% accuracy and 96.83% macro F1-score, outperforming RGB-only and sensor-only baselines. These results indicate that wearable motion cues provide robust complementary information when RGB evidence is degraded, and that calibrated decision-level fusion can further reduce modality-specific errors. Future work will extend the dataset, evaluate more diverse acquisition settings, and investigate signer-independent calibration and real-time deployment.

Data availability

The datasets generated and analysed during the current study are available from the corresponding author upon reasonable request. Public sharing of the raw RGB video data is restricted because the data contain potentially identifiable participant images. Processed sensor data and metadata required to reproduce the main analyses may be made available for academic research purposes, subject to institutional approval and participant consent restrictions.

Code availability

The custom code used for model training, evaluation, temperature selection and decision-level fusion is available from the corresponding author upon reasonable request.

References

1. Zhang, J. *et al.* Sign language recognition based on dual-path background erasure convolutional neural network. *Sci. Reports* **14**, 11360 (2024).

2. Qin, J. & Wang, M. Sign language recognition based on dual-channel star-attention convolutional neural network. *Sci. Reports* **15**, 27685 (2025).
3. Aly, M. & Fathi, I. S. Recognizing american sign language gestures efficiently and accurately using a hybrid transformer model. *Sci. Reports* **15**, 20253 (2025).
4. Faisal, M. A. A., Abir, F. F., Ahmed, M. U. & Ahad, M. A. R. Exploiting domain transformation and deep learning for hand gesture recognition using a low-cost dataglove. *Sci. Reports* **12**, 21446 (2022).
5. Rastgoo, R., Kiani, K. & Escalera, S. Sign language recognition: A deep survey. *Expert. Syst. with Appl.* **164**, 113794 (2021).
6. Li, D., Rodriguez, C., Yu, X. & Li, H. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1459–1469 (2020).
7. Sincan, O. M. & Keles, H. Y. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access* **8**, 181340–181355 (2020).
8. Camgoz, N. C., Koller, O., Hadfield, S. & Bowden, R. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10023–10033 (2020).
9. Duarte, A. *et al.* How2sign: A large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2735–2744 (2021).
10. Liu, Z. *et al.* Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3202–3211 (2022).
11. Arnab, A. *et al.* Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6836–6846 (2021).
12. Tong, Z., Song, Y., Wang, J. & Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Adv. Neural Inf. Process. Syst.* **35**, 10078–10093 (2022).
13. Li, Y. *et al.* Mvity2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4804–4814 (2022).
14. Hu, H., Zhao, W., Zhou, W., Wang, Y. & Li, H. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11087–11096 (2021).
15. Min, Y., Hao, A., Chai, X. & Chen, X. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11542–11551 (2021).
16. Zuo, R. & Mak, B. C2slr: Consistency-enhanced continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5131–5140 (2022).
17. Hu, L., Gao, L., Liu, Z. & Feng, W. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2529–2539 (2023).
18. Zheng, J. *et al.* Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23141–23150 (2023).
19. Jiang, S. *et al.* Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 3413–3423 (2021).
20. Kudrinko, K., Flavin, E., Zhu, X. & Li, Q. Wearable sensor-based sign language recognition: A comprehensive review. *IEEE Rev. Biomed. Eng.* **14**, 82–97 (2021).
21. Lee, B. G., Chong, T. W. & Chung, W. Y. Sensor fusion of motion-based sign language interpretation with deep learning. *Sensors* **20**, 6256 (2020).
22. Saggio, G. *et al.* Sign language recognition using wearable electronics: Implementing k-nearest neighbors with dynamic time warping and convolutional neural network algorithms. *Sensors* **20**, 3879 (2020).
23. Ji, A. *et al.* Dataglove for sign language recognition of people with hearing and speech impairment via wearable inertial sensors. *Sensors* **23**, 6693 (2023).
24. Wen, F., Zhang, Z., He, T. & Lee, C. Ai enabled sign language recognition and vr space bidirectional communication using triboelectric smart glove. *Nat. Commun.* **12**, 5378 (2021).

25. DelPreto, J., Hughes, J., D’Aria, M., de Fazio, M. & Rus, D. A wearable smart glove and its application of pose and gesture detection to sign language classification. *IEEE Robotics Autom. Lett.* **7**, 10589–10596 (2022).
26. Yu, J., Qin, M. & Zhou, S. Dynamic gesture recognition based on 2d convolutional neural network and feature fusion. *Sci. Reports* **12**, 4345 (2022).
27. Shin, J. *et al.* Hand gesture recognition using semg signals with a multi-stream time-varying feature enhancement approach. *Sci. Reports* **14**, 22061 (2024).
28. Tacca, N. *et al.* Wearable high-density emg sleeve for complex hand gesture classification and continuous joint angle estimation. *Sci. Reports* **14**, 18564 (2024).
29. Maashi, M., Iskandar, H. G. & Rizwanullah, M. Iot-driven smart assistive communication system for the hearing impaired with hybrid deep learning models for sign language recognition. *Sci. Reports* **15**, 6192 (2025).
30. Moin, A. *et al.* A wearable biosensing system with in-sensor adaptive machine learning for hand gesture recognition. *Nat. Electron.* **4**, 54–63 (2021).
31. Yuan, G. *et al.* Hand gesture recognition using deep feature fusion network based on wearable sensors. *IEEE Sensors J.* **21**, 539–547 (2021).
32. Li, W., Luo, Z., Li, W. & Xi, X. Chinese sign language recognition based on surface electromyography and motion information. *PLOS ONE* **18**, e0295398 (2023).
33. Lu, C., Amino, S. & Jing, L. Data glove with bending sensor and inertial sensor based on weighted dtw fusion for sign language recognition. *Electronics* **12**, 613 (2023).
34. Kothadiya, D. R., Bhatt, C. M., Kharwa, H. & Albu, F. Hybrid inceptionnet based enhanced architecture for isolated sign language recognition. *IEEE Access* **12**, 90889–90899 (2024).
35. Ozdemir, O., Baytas, I. M. & Akarun, L. Multi-cue temporal modeling for skeleton-based sign language recognition. *Front. Neurosci.* **17**, 1148191 (2023).
36. Bohacek, M. & Hruz, M. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 182–191 (2022).
37. Pradhan, A., He, J. & Jiang, N. Multi-day dataset of forearm and wrist electromyogram for hand gesture recognition and biometrics. *Sci. Data* **9**, 733 (2022).
38. Dinh, N. S. *et al.* Sign language recognition: A large-scale multi-view dataset and comprehensive evaluation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1211–1221 (2025).
39. Shin, J., Miah, A. S. M., Suzuki, K., Hirooka, K. & Hasan, M. A. M. Dynamic korean sign language recognition using pose estimation based and attention-based neural network. *IEEE Access* **11**, 143501–143513 (2023).
40. Pal, D., Kumar, A., Kumar, V., Basangar, S. & Tomar, P. Development of an otdr-based hand glove optical sensor for sign language prediction. *IEEE Sensors J.* **24**, 2807–2814 (2024).
41. Sumbul, H. A novel mems and flex sensor-based hand gesture recognition and regenerating system using deep learning model. *IEEE Access* **12**, 133685–133693 (2024).
42. Chen, G. *et al.* A novel illumination-robust hand gesture recognition system with event-based neuromorphic vision sensor. *IEEE Transactions on Autom. Sci. Eng.* **18**, 508–520 (2021).
43. Fiorini, L. *et al.* Daily gesture recognition during human–robot interaction combining vision and wearable systems. *IEEE Sensors J.* **21**, 23568–23577 (2021).
44. Liu, Y. *et al.* A wearable system for sign language recognition enabled by a convolutional neural network. *Nano Energy* **116**, 108767 (2023).
45. Minderer, M. *et al.* Revisiting the calibration of modern neural networks. *Adv. Neural Inf. Process. Syst.* **34**, 15682–15694 (2021).
46. Balanya, S. A., Maronas, J. & Ramos, D. Adaptive temperature scaling for robust calibration of deep neural networks. *Neural Comput. Appl.* **36**, 8073–8095 (2024).

Acknowledgements and Funding

This work was supported in part by the Science and Technology Correspondent Project of Tianjin under Grant 25YDTPJC00380.

Author contributions

X.H. conceived the study, designed the RWS-Fusion framework, developed the RGB–wearable sensor fusion method, conducted the experiments, analysed the results and drafted the manuscript. C.H. contributed to the development of the wearable sensing system, data acquisition, experimental validation and result analysis. W.W. supervised the project, provided methodological guidance and revised the manuscript. T.X. contributed to system construction, data collection, data preprocessing, experimental assistance and manuscript revision. Z.L. contributed to wearable-system implementation, technical support, data preprocessing, experimental assistance and manuscript revision. C.Z., Y.S. and J.Z. contributed to system construction, data collection, data preprocessing, experimental assistance and manuscript revision. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.